Implementación de un sistema de captura de gestos usando un leap motion y redes neuronales para su clasificación.

Gabriel Isaac Muñoz Garcia

Implementación de un sistema de captura de gestos usando un leap motion y redes neuronales para su clasificación.

Gabriel Isaac Muñoz Garcia

Trabajo de grado presentado como requisito
parcial para optar al título de
Ingeniero de sistemas y computación

Pereira, Septiembre de 2017
UNIVERSIDAD TECNOLÓGICA DE PEREIRA
Programa de Ingeniería en Sistemas y computación
Grupo de investigación SIRIUS

Implementación de un sistema de captura de gestos usando un leap motion y redes neuronales para su clasificación.

Gabriel Isaac Muñoz Garcia

Director: John Haiber Osorio

Pereira, Septiembre de 2017
Programa de Ingeniería en Sistemas y computación.
Universidad Tecnológica de Pereira
La Julita. Pereira(Colombia)
TEL: (+57)(6)3137122
www.utp.edu.co

Capítulo 1

Introducción

Se describe a continuación el proceso llevado a cabo para la implementación de un sistema de captura de gestos usando un sensor propietario y redes neuronales, el prototipo se construyó usando tecnologías web y algunos frameworks para javascript junto con un sensor llamado "leap motion" [1] usado como dispositivo de entrada. Los datos otorgados por el sensor son la base para el entrenamiento de las redes neuronales a usar. En las pruebas se compararon los 2 modelos; clasificación multi-etiqueta(ML) y clasificación multi-clase(MC), teniendo en cuenta factores como: tiempos de entrenamiento, error aproximado y cantidad de datos.

1.1. Descripción del proyecto

Se creó un prototipo funcional el cual permite por medio de una interfaz, realizar la captura de los datos para el entrenamiento del sistema de una manera sencilla ofreciendo libertad al usuario de crear gestos predeterminados. Durante el proceso se evaluaron las capacidades del Leap motion para reconocer las falanges de las manos en diferentes ocasiones y la eficiencia de las técnicas clasificación del aprendizaje de máquinas usadas para reconocer gestos del alfabeto dactilológico de la lengua de señas colombiano.

Por medio de las pruebas se pretende comparar 2 métodos de clasificación para varias clases e identificar el más apropiado para una interfaz, midiendo diferentes aspectos entre ellos como: su tiempo de entrenamiento y cantidad de datos necesarios para este, también se creó para el prototipo una serie de pantallas para visualizar la información a lo largo del proceso.

Capítulo 2

Marco Teórico

2.1. Movimiento Humano

Para entender un poco más se debe hablar de los orígenes del estudio del movimiento humano y la aparición de tecnologías para automatizar dichos procesos, el estudio del movimiento humano está basado en analizar patrones de mociones globales en vez de patrones locales tales como gestos de las manos o expresiones faciales.

El interés en el movimiento humano es motivado por la curiosidad, necesidad o métodos disponibles y también se puede ver en campos como las matemáticas o las artes para la descripción de fenómenos visuales. Todas estas formas de moción fueron presentadas siempre de forma estática y solo empezó a hablarse de representaciones dinámicas de movimiento a finales del siglo XIX, estos patrones de movimiento humano eran usualmente estudiados en relación cercana a mociones en patrones animales. Ciertamente, estos estudios comparativos siguen siendo útiles. En el siglo XX, la biomecánica se convirtió en una disciplina que permitió establecer estándares para la ingeniería humana en Europa y Estados Unidos, trayendo tiempos después consigo tecnologías de análisis cinematográfico que permiten realizar el estudio detallado de una actividad física. Con esto empezaron a surgir estudios importantes entre ellos la concepción espacial de problemas como: "los grados de libertad en actividades físicas (correr", caminar o saltar) y "la eficiencia y el costo de energía en el movimiento humano". Estos estudios permitieron dar paso a la biofísica como puede verse en [2].

El estudio de la captura del movimiento humano es realizado por equipos avanzados capaces de captar traslación y rotación de puntos de interés y digitalizarlos para su estudio, el proceso básicamente consiste en modelar una forma humana(dinámica) aplicando perspectiva geométrica, esta es la base del 3D y proviene de la época del renacimiento (luces y sombras), junto los estudios más adelante sobre los rayos visuales en la geometría óptica que se encarga de estudiar la propagación de la luz en términos de rayos.

2.2. Captura de movimiento (MOCAP)

Existen muchas formas de realizar captura de movimiento por ello la forma de recopilar los datos varía según el formato y el tipo de tecnología usada para la recolección de estos, existen desde captura de movimientos por medio de marcadores y cámaras de alta velocidad que son sensibles a colores, hasta dispositivos que capturan los movimientos por medio de un chequeo constante de los campos magnéticos que emiten unos transmisores en un espacio dado. Estas tecnologías se les categoriza según su forma de integrarse y su uso a la hora de recolectar datos y pueden ser invasivos (wearables) o no invasivos. Se hablará un poco más sobre las tecnologías usadas para la captura de movimientos y la estructura de los datos existentes a la hora de realizar una captura de un movimiento de una persona u objeto.

Los sistemas de MOCAP generan una enorme cantidad de datos y estos provienen de las posiciones y rotaciones de cada una de las acciones que hacen parte de los movimientos que se pretenden grabar, luego de obtener los datos en el formato deseado es posible realizar diferentes tipos de tareas corriendo un serie de análisis sobre los datos según la necesidad, como por ejemplo: realizar análisis biomecánicos, animaciones o servir como dispositivos para interactuar en tiempo real con objetos virtuales.

Desde tiempos antiguos de la formación de este arte, los animadores han observado el movimiento real de criaturas en orden de crear un movimiento animado y tener la capacidad de manipular el movimiento de personajes u objetos, esto da una libertad a el animador pero a su vez se convierte en una maldición debido a que debe tener bajo control muchas partes involucradas en el movimiento, y a frecuentemente esto representa una tarea que requiere trabajo y habilidad, lo mismo sucede con los datos.

A menudo los animadores usan estos movimientos grabados para transferirlos a personajes animados o modelos tridimensionales a esto se le conoce como çaptura de movimiento", incluso sabiendo que solo representa uno de los aspectos para crear animación desde la observación de un movimiento real.

La captura de movimiento crea una representación la cual difiere movimiento a partir de la apariencia; eso codifica el movimiento en una forma que es adecuada para las clases de procesamiento o análisis que se necesitan realizar.[3]

Existen sistemas de captura de tiempo real en donde se requiere que la animación se produzca instantáneamente, a esto se le conoce como "digital puppeteering", también existe captura de movimientos parcial o de cuerpo completo.

La captura de movimiento puede ser realizada por una cantidad de razones diferentes a la animación como se mencionó anteriormente, por ejemplo en cosas como análisis biomédicos, análisis en desempeño de deportes o inclusive usarse como entrada de un sistema de interacción humano-computador, pero estas actividades no difieren mucho debido a que en sus etapas tempranas se necesitan crear observaciones que van a ser interpretadas.

Existen una variedad de métodos para capturar mociones a cierto nivel y la tecnología actual para leer y guardar un movimiento es irrelevante debido a que estas llegaran a el mismo resultado si tienen formatos en común, pero la calidad y fluidez de las animaciones

dependerá de la tecnología usada. Sin embargo cada aproximación tiene sus ventajas y desventajas y difieren en que la experiencia de cada usuario es limitada por las características del dispositivo.

Entre las tecnologías usadas actualmente en el campo del MOCAP existen las magnéticas y las ópticas, estas tecnologías y sistemas están diseñados para rastrear mociones de figuras humanas y requieren un esqueleto mecánico al cual se deberán enlazar, hoy en día las implementaciones modernas utilizan mecanismos avanzados para reducir la carga.

Las tecnologías magnéticas de captura usan herramientas que establecen campos magnéticos en un espacio y luego usan sensores que permiten determinar la posición y orientación en el espacio basado en estos campos, cada marcador o sensor tiene su propio canal de datos, pero son sensibles a cualquier objeto metálico, este ha sido una opción en el campo de la animación debido a el control de flujo de información que posee gracias a sus canales.

Los sistemas de rastreo óptico usan marcadores visuales especiales y un número especial de cámaras de alta velocidad para determinar la posición del objeto, estos marcadores generalmente son objetos pasivos como esferas retro-reflectivas o dispositivos monocromáticos ajustados para captar un color específico de luz, estos marcadores se pueden perder debido a la oclusión y a menudo se usa un gran número de cámaras para compensar y reducir ese riesgo, una de las mayores desventajas de estos métodos es que a pesar de que estos sistemas pueden ver el objeto, no tienen una forma de diferenciar qué marcador es cual y a diferencia de los sistemas magnéticos. (algunos algoritmos para procesamiento de imágenes digitales están cambiando esto actualmente).

Estos dispositivos encargados del análisis de marcha en humanos por medio de marcadores brillantes suelen asociarse a un modelo tridimensional para grabar el movimiento, estos datos luego son tomados generalmente de las uniones entre las extremidades y permiten tener una forma de reproducir de nuevo el movimiento digitalmente, existen muchos otros dispositivos que permiten recopilar estos datos. Estos se pueden diferenciar en invasivos y no invasivos teniendo en cuenta el grado de conexión con el cuerpo, entendiendo como invasivo un objeto que debe ser usado en el cuerpo, por lo general los no invasivos no necesitan cámaras de tan altas velocidades gracias a el avance de la tecnología, por otro lado existen también aquellos que funcionan por medio de campos magnéticos y no usan cámaras en lo absoluto.

Debido a los problemas de oclusión y correspondencia de los marcadores los sistemas magnéticos, por lo general su costo suele ser menor, es importante resaltar que ambas tecnologías están cambiando rápidamente compensando muchos de los problemas que existen.

La captura de datos de movimientos es imperfecta y existen algunos aspectos que se deben tener en cuenta al tratar los datos y estos son cosas como:

- Reutilización La captura de datos graba un evento, si se quieren usar estos datos para un fin distinto o una acción diferente se necesita editar los datos o darles un formato a estos.
- Imperfección de la realidad

Los movimientos reales no son perfectos, no siempre se realiza una acción de la misma forma y cuando es un movimiento repetitivo no necesariamente este es cíclico.

- Cambio de propósito

  No es posible predecir qué movimiento se necesitara y si se predice alguien o algo cambiara su mente frente a lo que deseaba.

  Rastrear movimiento humano ha sido un tópico importante en la visión computacional, de todos modos para la mayoría de aplicaciones crear una representación en 3d de el movimiento no es requerido. Por ejemplo crear una representación 2D del movimiento para identificación de acciones y una gran variedad de demostraciones interactivas están basadas en "silhouette tracking", aplicaciones como interfaces requieren rendimiento a veces a costo de fidelidad(performarnce-liability).

2.3. Inteligencia artificial y machine learning

Se entiende por inteligencia a la capacidad de aprender y adecuarse a ciertos problemas gracias a la facultad de la mente que permite aprender, razonar, tomar decisiones y formarse una idea determinada de la realidad. Mientras que la inteligencia artificial es una área multidisciplinaria que a través de las ciencias de la computación, matemática, lógica y la filosofía, estudia la creación y diseño de sistemas capaces de resolver problemas cotidianos por sí mismas usando como paradigma la inteligencia humana.[4]

Según John Mccarthy quien en 1956 definió la inteligencia artificial como "La ciencia e ingenio de hacer máquinas inteligentes especialmente programas de cómputo inteligentes."

Está representada por un conjunto de métodos como lo son: Árboles de decisiones, Algoritmos genéticos.(análogo al proceso de evolución del ADN), Redes neuronales artificiales(análogo al funcionamiento del cerebro.), Razonamiento bajo lógicas formales, SVM (Support Vector Machines.)[5] y muchos otros métodos que existen para realizar clasificación sobre diferentes tipos de datos

Estos métodos son estudiados en el campo del "machine learning.º aprendizaje de máquina y han surgido gracias a el estudio en diferente ámbitos, una gran parte es usada para encontrar patrones de reconocimiento y avanzar en la teoría de aprendizaje computacional en inteligencia artificial, además busca explorar la construcción de algoritmos que pueden aprender y hacer predicciones a partir de datos, por lo general estos algoritmos funcionan usando ejemplos como entrada en orden de realizar predicciones y decisiones a partir de datos con un mismo formato en el futuro.

Existen muchas aplicaciones para el aprendizaje automático y sus diferentes métodos, ya que pueden ser usados en diferentes escenarios debido a su facultad para predecir o clasificar gran cantidad de datos con un tasa de fiabilidad que puede tocar márgenes muy altas, algunos de sus usos pueden verse en cosas como:

- A facebook profiled based TV recommender System: Donde hacen uso del algoritmo K-means y redes bayesianas y otros métodos para encontrar información en un grupo de personas en facebook y recomendar programas de acuerdo a las afinidades de los grupos.
- A novel system for hand gesture Recognition: Aquí un sistema de captura de gestos básicos donde se usa un guante especial. Estos movimientos se interpretan por la aplicación con la comodidad de una cámara web. Luego estas mociones son analizadas usando computación visual y aprendizaje de máquina en este caso particular decidieron implementar HMM hidden markov models.
- Applying Reinforcement learning to competitive Tetris: Un intento de aplicación de aprendizaje reforzado en el juego tetris el cual es jugado en un tablero de 20 filas y 10 columnas, se modela el problema también hacen uso de HMM (hidden markov models) debido a las transiciones probabilísticas de estado a estado.

2.4. Redes Neuronales

Las redes neuronales son métodos de aprendizaje automático, capaces de resolver problemas de clasificación o predicción realizando un ajuste de parámetros el cual está basado en un entrenamiento que se logra haciendo uso de datos muestra, también conocidos como set de datos. Estas muestras ayudarán en el proceso de entrenamiento del algoritmo y por lo general funcionan con el gradiente descendiente, lo cual permite minimizar el error que existe entre las muestras de acuerdo a la hipótesis planteada, logrando así realizar la clasificación o predicción según su ajuste de parámetros internos, es importante recordar que depende en gran parte de los datos provistos a el sistema.

Las redes neuronales requieren muchas veces y dependiendo de su configuración más o menos tiempo en entrenarse de acuerdo a la calidad, cantidad de datos y el numero de características que se pretenden analizar. El entrenamiento debe realizarse hasta un punto estable sin hacer que aparezca el problema de overfitting, el cual puede ocurrir debido a un sobre-entrenamiento y se pierde la habilidad de la red para generalizar, es vital recordar que es un proceso lento, el cual lleva mucho más tiempo a medida que el número de datos crece.

La configuración de una red neuronal permite ajustar parámetros como tasas de aprendizaje, capas ocultas, número de neuronas, iteraciones, rango de error y además ajustar cada cuantas iteraciones se mostrará información sobre el entrenamiento de la red.

Capítulo 3

Metodología

A continuación se describen las tareas llevadas a cabo durante las diferentes fases técnicas del proyecto para alcanzar los objetivos propuestos y documentar los resultados del desarrollo.

Con el fin de investigar un poco más a fondo el caso de estudio "la lengua de señas" se decidió acudir a instituciones relacionadas con el tema, entre ellas: La escuela de la palabra, fundación conectando sentidos y tecnoparque. Las cuales promueven soluciones para los sectores con estos problemas, además facilitaron el acceso a información pertinente al tema como los datos estadísticos de la población actual con discapacidad de algún tipo.

Al conocer sobre las necesidades de las personas con discapacidad auditiva y explorar un poco sobre el uso de esta lengua y los sistemas existentes para su interpretación en un entorno virtual, la implementación fue guiada por la necesidad de ajustar la herramienta al problema del desconocimiento de la lengua de señas en alguno de sus escenarios, pero durante el proceso se pudo observar que es fundamental diseñar la herramienta teniendo en cuenta las necesidades de los usuarios, por lo tanto se decidió crear una prueba de concepto usando las configuraciones manuales usadas en el alfabeto de la lengua de señas dactilológico colombiano como caso de estudio para este sistema.

Se recolectó información de apoyo para aprender sobre métodos de aprendizaje de máquina y detalles de la implementación de algunos de sus métodos. Luego se diseñó un borrador para la prueba de los métodos en javascript, ayudado de diagramas. Tras encontrar frameworks y librerías que permiten realizar la comparación, se usó una plantilla para dar un estilo agradable al entorno.

Durante la etapa de selección se escogieron las tecnologías a usar para la creación del prototipo. Entre ellas un sensor de captura de movimiento para la recolección de datos de entrenamiento (Leap motion) y un ambiente de desarrollo creado en la web para representar los diferentes pasos a seguir dentro del entrenamiento de una red neuronal:

- Toma de datos de entrenamiento.
- Agrupación de datos según muestras.

- Etiquetado de los grupos, entrenamiento de la red.
- Guardado de las redes neuronales con su respectivos pesos sinápticos luego del entrenamiento.

Para la implementación se usaron algunos frameworks de javascript como angular.js[6], synaptic.js[7] y leap.js[8], entre otros. En el backend o manejo de los datos se usó una base de datos no relacional creada en mongoDB[9].

Con el fin de estructurar la aplicación y sus diferentes módulos se decidió seguir una metodología de desarrollo ágil, en este caso kanban[10], para las tareas de desarrollo y de diseño, tratando de ajustarse al cronograma propuesto en la ficha técnica, la arquitectura empleada fue:

Una arquitectura cliente servidor, con un backend en django y permanencia de datos gestionada con el motor de base de datos MongoDB, acompañado con un front-end usando html, js, css junto con algunos frameworks de javascript mencionados anteriormente.

Luego de crear un prototipo funcional que mostrara el concepto se decidió crear el documento con los resultados durante el proceso de implementación del prototipo final para ajustarse a objetivos generales y específicos del proyecto.

Se creo el esquema a seguir para realizar las pruebas y recolectar datos con el fin de soportar lo encontrado durante el desarrollo del proyecto, las pruebas consistieron en entrenar diferentes modelos de redes neuronales con los mismos datos y determinar cual de ellos era más adecuado para una interfaz en la web, evaluando principalmente el tiempo de entrenamiento a medida que aumenta el número de gestos a clasificar.

Por medio de las pruebas se compararon tiempos de entrenamiento de los métodos entre otros aspectos como cantidad de datos y error aproximado entre las muestras, luego se crearon métodos de visualización para los datos recolectados para ayudar en las diferentes etapas del proceso.

Capítulo 4

Desarrollo

Durante la Investigación y documentación sobre el sistema creado “gesture control”, se usaron diferentes recursos tales como: artículos, libros y tutoriales sobre la implementación de diferentes arquitecturas de redes neuronales y su funcionamiento, luego se decidió realizar el diseño de un diagrama de bloques seguido de la descomposición del trabajo.

Una vez seleccionadas las herramientas se diseñó un esquema con los componentes involucrados para alcanzar los objetivos propuestos y los módulos necesarios. La aplicación se dividió en varios componentes: persistencia de datos, plantilla, contenido externo, archivos estáticos y página web la cual consta de diferentes módulos para captura de datos, visualización, etiquetado, agrupación, entrenamiento y pruebas.

Se proveerá una descripción breve de lo que contiene cada uno de estos módulos y su papel en el uso de la aplicación por medio de un diagrama de flujo de los componentes involucrados.

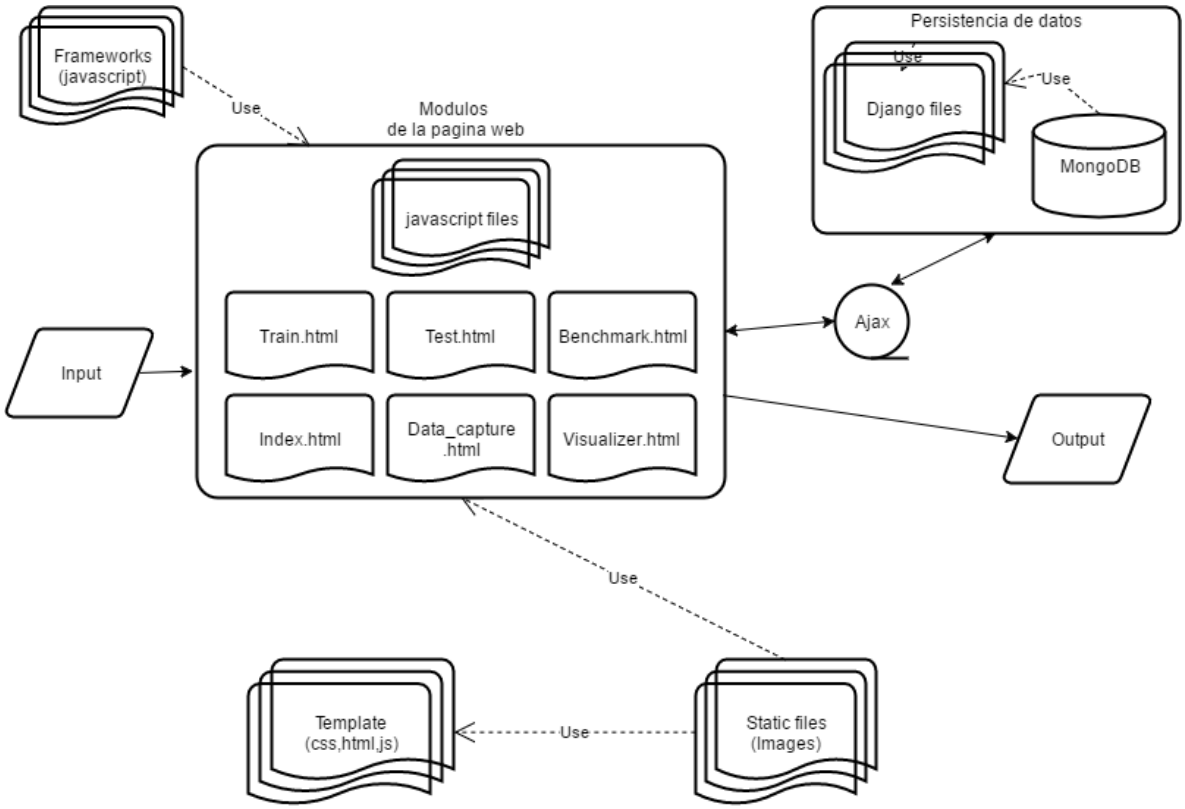

Figura 4.1: Diagrama de módulos: muestra la asignación de clases y objetos o módulos en el diseño del sistema

4.1. Home

Pantalla inicial con tema de la aplicación.

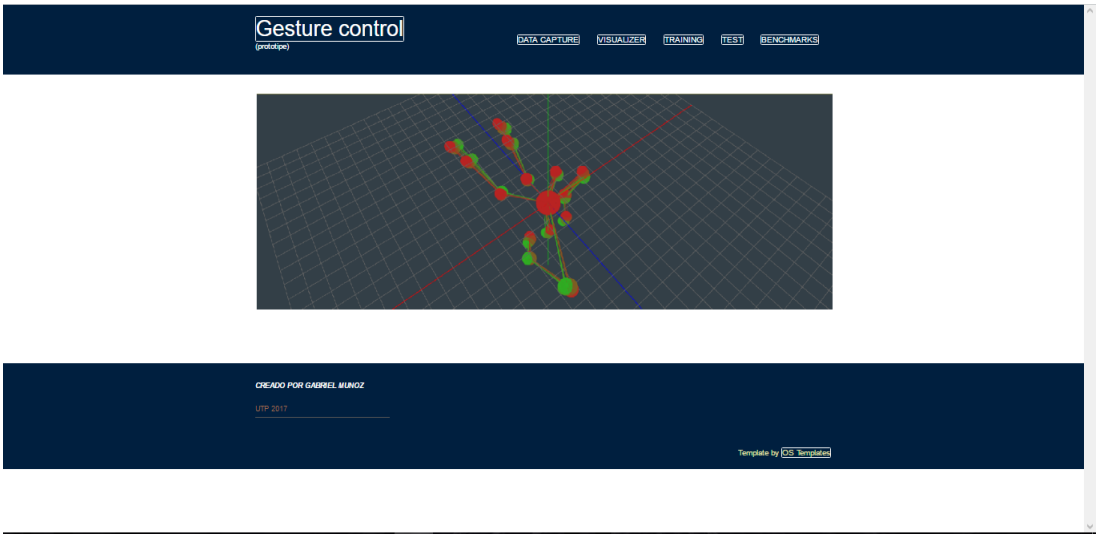

Figura 4.2: Home: contiene menú de navegación de la aplicación, permite redirigir a cualquier modulo de esta.

4.2. Captura de datos

El prototipo posee una interfaz capaz de capturar un número determinado de muestras de la posición de las falanges de las manos y asociarlos a un nombre para luego visualizarlas usando three.js[11]. Como indicador visual del progreso durante la toma de datos, se creó una barra de carga que al terminar este proceso realiza una pre-visualización de los datos tomados en 3D para descartarlos o guardarlos para pasos posteriores.

Figura 4.3: pantalla de captura de datos: usada para agregar el nombre del gesto a grabar (primer campo) y el numero de muestras de este(segundo campo).

4.3. Visualización

Esta interfaz permite al usuario:

- Ver todas las muestras tomadas de los gestos a usar (cargarlas de la base de datos y mostrarlas por pantalla)
- Filtrar por nombre las muestras.
- Crear un set de datos a partir de estas muestras y agruparlas.
- Permite que los sets de datos sean guardados.

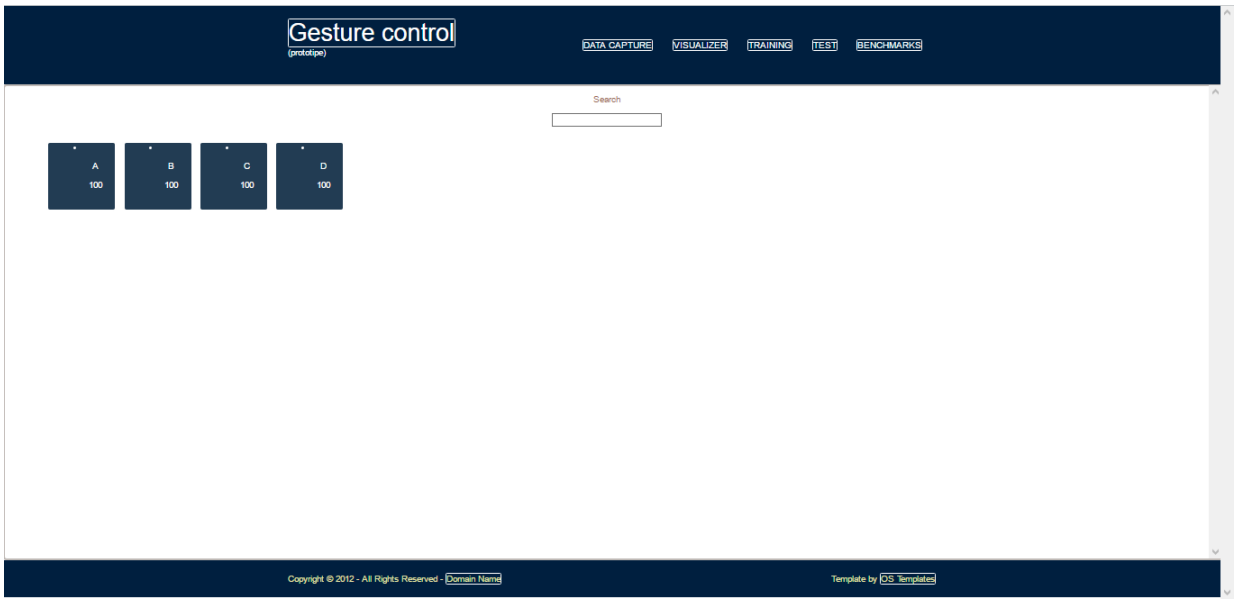

Figura 4.4: Pantalla de visualización: usada para ver la muestra seleccionada en un entorno 3D.

4.4. Agrupación

Se seleccionan los gestos a ser parte del set de datos durante la etapa de entrenamiento y se miran las proporciones y cantidad de datos por gesto. Se creó una gráfica dinámica para mantener un equilibrio al seleccionar muestras y no tener problemas a la hora del entrenamiento de la red, procurando tener un número de datos consistente para cada gesto.

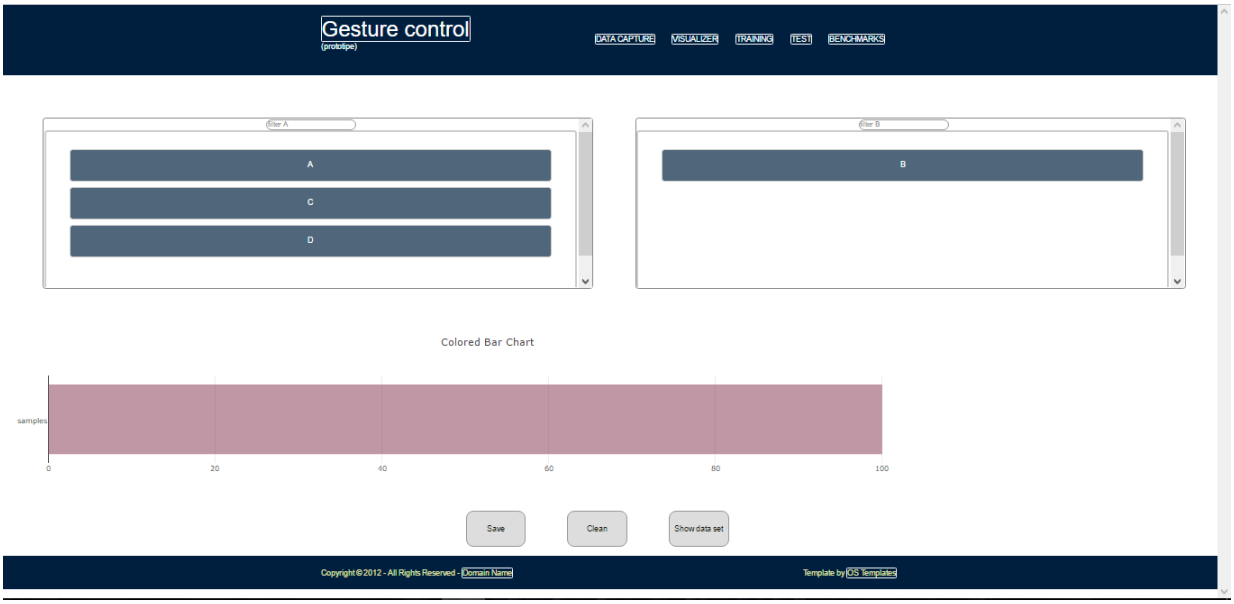

Figura 4.5: Pantalla de agrupación de muestras. Usada para arrastrar las muestras requeridas a la derecha para la creación del set de datos.

4.5. Etiquetado

Se provee formato al set de datos seleccionado para el entrenamiento con sus respectivas etiquetas, las cuales permite asociar el conjunto de muestras a una clase y proveer un indicador visual durante el proceso de entrenamiento y pruebas.

4.6. Entrenamiento

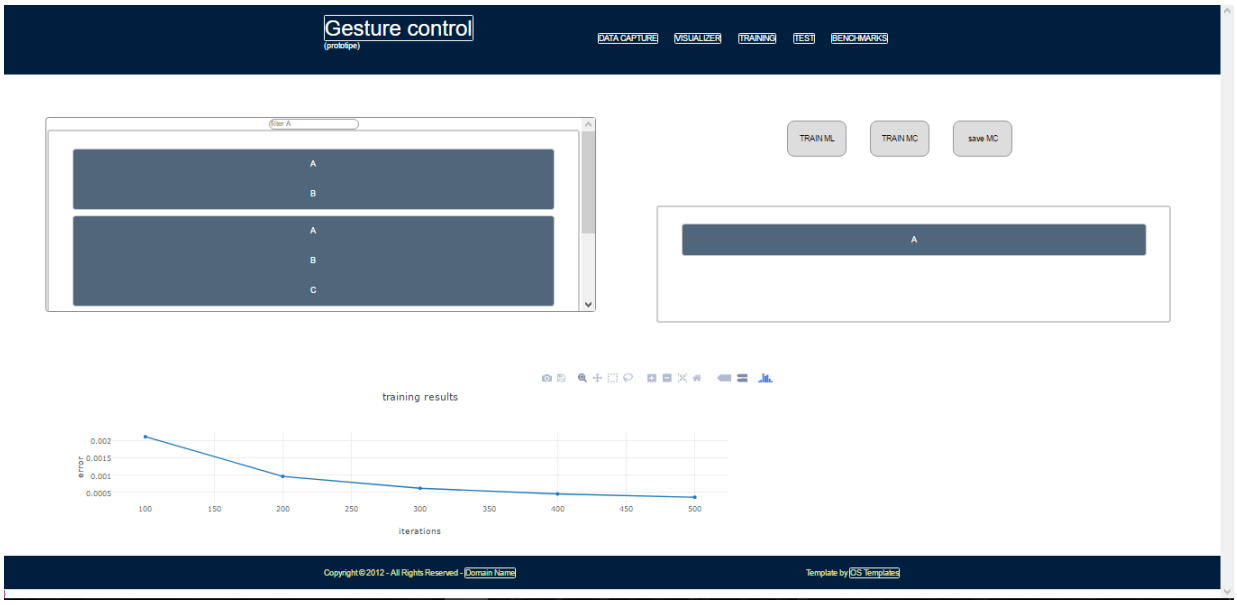

Figura 4.6: Pantalla de entrenamiento: Usada para seleccionar el set de datos y red a entrenar una vez es arrastrado al campo de la derecha.

Se selecciona el set de datos y se elige el tipo de red a entrenar (el tiempo que lleva realizar el entrenamiento de la red, puede variar dependiendo de los parámetros seleccionados entre 5min a 60min máximo para mantener márgenes prudentes de tiempo sin bloquear el navegador debido al entrenamiento síncrono), al terminar el entrenamiento se muestran gráficas del error aproximado obtenido con las muestras usadas en el modelo seleccionado y se decide si se quiere guardar la red neuronal en el estado actual en la base de datos para cargarla y realizar las pruebas.

Se describe a continuación los parámetros usados para el entrenamiento de los 2 modelos estudiados y sus respectivas arquitecturas:

Arquitectura ML

- Layout: [45 neuronas,30 neuronas,30 neuronas,30 neuronas,30 neuronas, N neuronas].
- Completamente conectada.
- Tasa de aprendizaje: 0.01.
- Error mínimo para terminar el entrenamiento: 0.00005

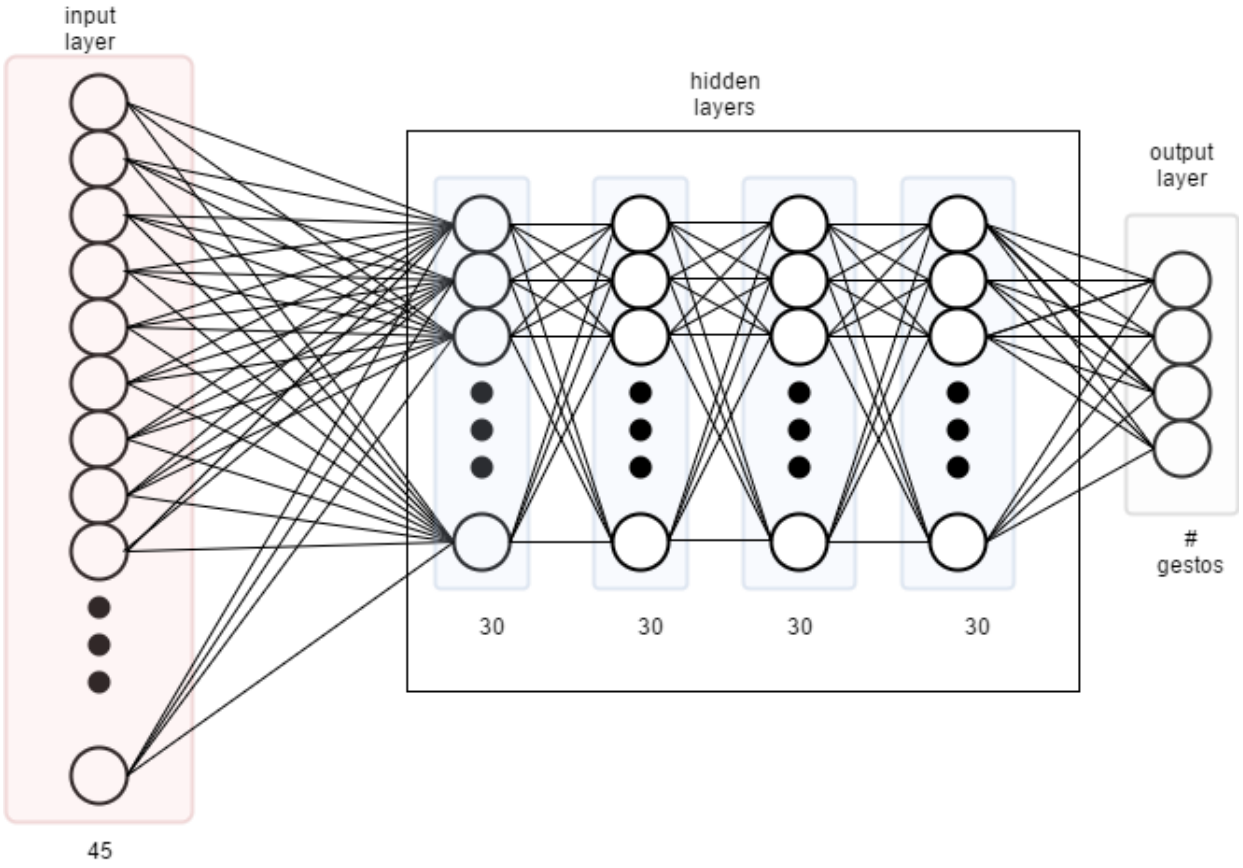

Figura 4.7: Modelo planteado para clasificación Multi-etiqueta(Busca clasificar los gestos en una sola estructura).

Arquitectura MC

- Diseño: [45 neuronas,15 neuronas, 5 neuronas,1 neurona].
- Completamente conectada.
- Tasa de aprendizaje: 0.01.
- Error mínimo para terminar el entrenamiento: 0.00005

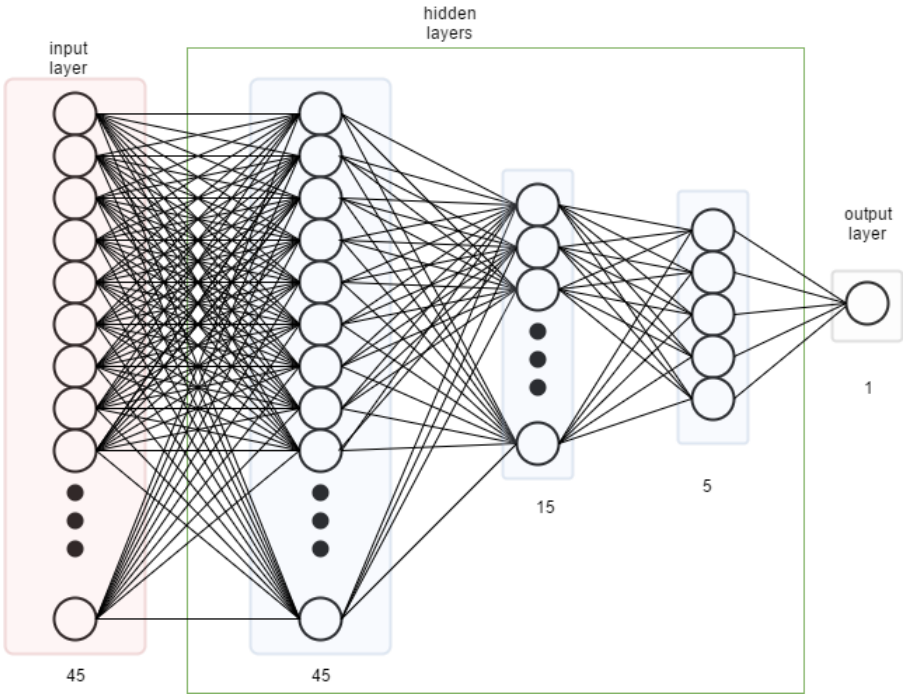

Figura 4.8: Modelo planteado para clasificación Multi-clase (Busca usar N redes de este tipo cada una encargada de la clasificación 1 solo gesto).

4.7. Pruebas

Se permite seleccionar las redes a entrenar y pasar los datos de prueba por medio del sensor nuevamente, se muestra el funcionamiento de la red seleccionada por medio de una gráfica y se realiza las mediciones de las salidas generadas para luego mostrarlas por pantalla como gráficas de barras para determinado gesto [0,1] (0 salida sin estímulo, 1 salida estimulada).

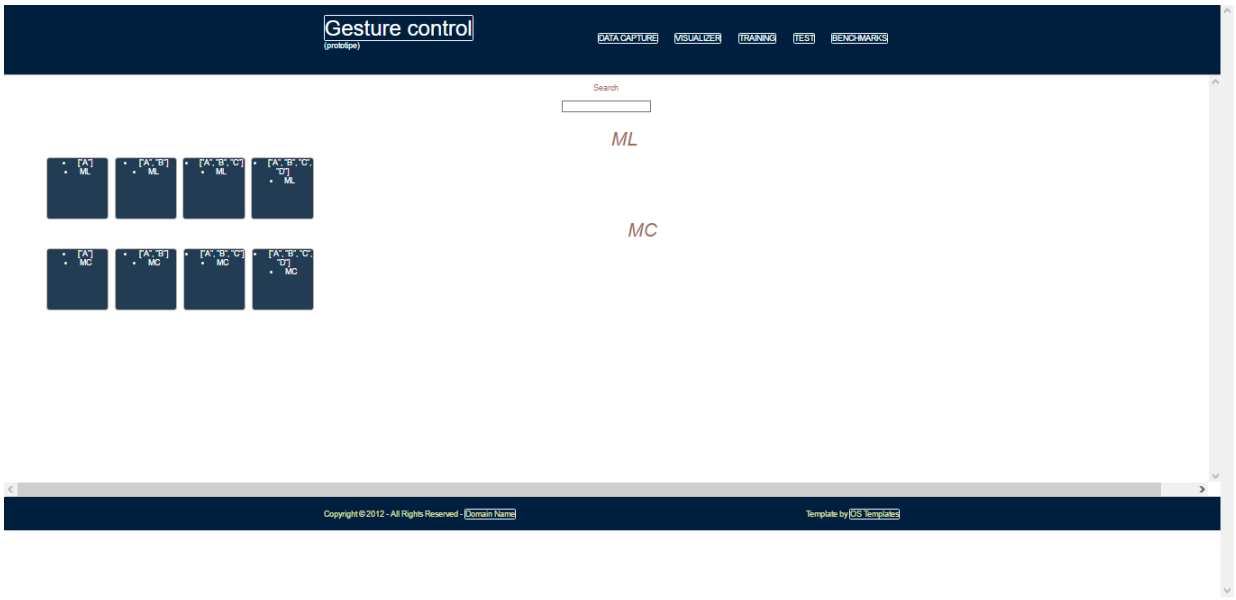

Figura 4.9: Pantalla de prueba para modelo. Usada para probar el desempeño de las redes entrenadas usando datos directamente del sensor.

Capítulo 5

Análisis de resultados

Este capítulo pretende presentar los resultados obtenidos durante el desarrollo del prototipo funcional para clasificación de gestos del alfabeto de señas dactilológico colombiano, para hacer el análisis se realizaron pruebas sobre 2 modelos, los cuales fueron:

- Una red neuronal para clasificación multi-etiqueta con **m** entradas y **n** salidas usando 4 capas de neuronas ocultas, buscando la salida que haya generado mayor estímulo con determinada muestra como salida de la red neuronal (ver fig 4.7).
- Una clasificación multi-clase usando un grupo de redes encargadas solo de clasificación **1 v N** con **m** entradas y **1** salida, luego se realiza una selección de acuerdo a la salida de la red mayor estimulada con la misma entrada. (ver fig 4.8).

Las pruebas se implementaron tomando datos sobre:

- La cantidad de gestos a clasificar.
- Cantidad de datos usados en entrenamiento.
- Tiempo.
- iteraciones.

Definido el esquema de las pruebas se crearon escenarios de prueba usando el mismo set de datos para entrenar ambos modelos, las pruebas consistieron en usar gestos del alfabeto dactilológico de la lengua de señas colombiana y agruparlos de a 4 gestos máximo. Los gestos usados fueron gestos estáticos y están limitados a los primeros 4 caracteres del alfabeto [A, B, C, D] debido a que los tiempos de entrenamiento y el tamaño del set de datos aumentan considerablemente a medida que crece el número de gestos a clasificar.

En el segundo modelo, se descompuso el problema de la clasificación multi-etiqueta para los gestos, esto dividió el problema de clasificar varios gestos a la vez a una serie de redes

encargadas de distinguir 1 solo gesto y alimentarse a la vez con la misma entrada, luego se selecciona la mayor estimulada y así se determina la salida, esta estrategia involucro entrenar un clasificador por clase usando sus muestras como positivas y las de las otras clases a comparar como negativas, buscando encontrar un valor de confidencia sobre el cual se va a tomar la decisión, similar al modelo basado en etiquetas. Tomar decisiones significa aplicar en todos los clasificadores la misma muestra y predecir cuál es el resultado, luego reportar el valor con mayor confidencia.

Esta forma de resolver el problema a pesar de ser popular, es una heurística que resulto tener varios problemas mencionados en la literatura:

- La confidencia puede diferir entre los clasificadores (errores distintos para cada red entrenada).
- Inclusive si una distribución de las muestras de las clases es balanceada en el set de entrenamiento, los aprendices de las clasificaciones binarias se ven des-balanceados debido a la distribución atípica del set de datos negativo, los cuales suelen ser mucho mayores que los de la parte positiva.

5.1. Gráficas de entrenamientos

Las siguientes gráficas muestran el MSE(mean squeared error) encontrado para cada set de datos y numero de iteraciones necesarias para su convergencia.

Entrenamiento multi-etiqueta

Figura 5.1: Entrenamiento multi-etiqueta 1 gesto: Se puede ver como converge a error aproximado prudente en las iteraciones previstas

Figura 5.2: Entrenamiento multi-etiqueta 2 gestos: Se puede ver como toma más iteraciones para converger y el error es mucho más grande

Figura 5.3: Entrenamiento multi-etiqueta 3 gestos: Se puede ver como toma mucho más tiempo para encontrar una solución la red y el error al iniciar es mucho mayor.

Figura 5.4: Entrenamiento multi-etiqueta 4 gestos: Se puede ver como la red encuentra una solución finalizando las iteraciones y que el error entre las muestras es mucho mayor al iniciar el entrenamiento

Entrenamiento multi-clase

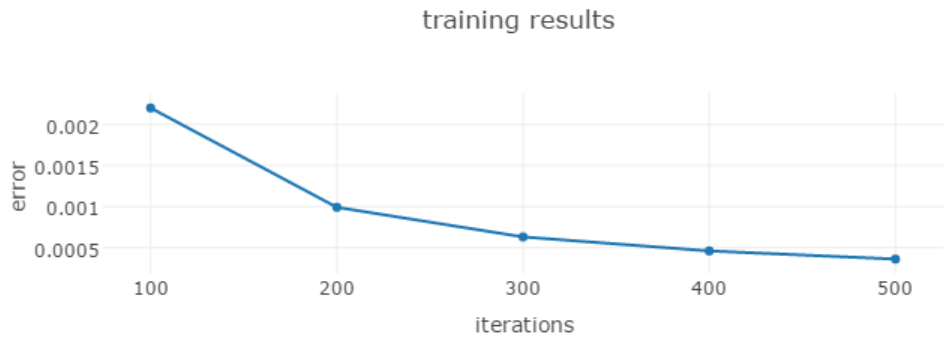

Figura 5.5: Entrenamiento multi-clase 1 gesto: Se puede ver que la gráfica se comporta similar a la del modelo multi-etiqueta para 1 solo gesto

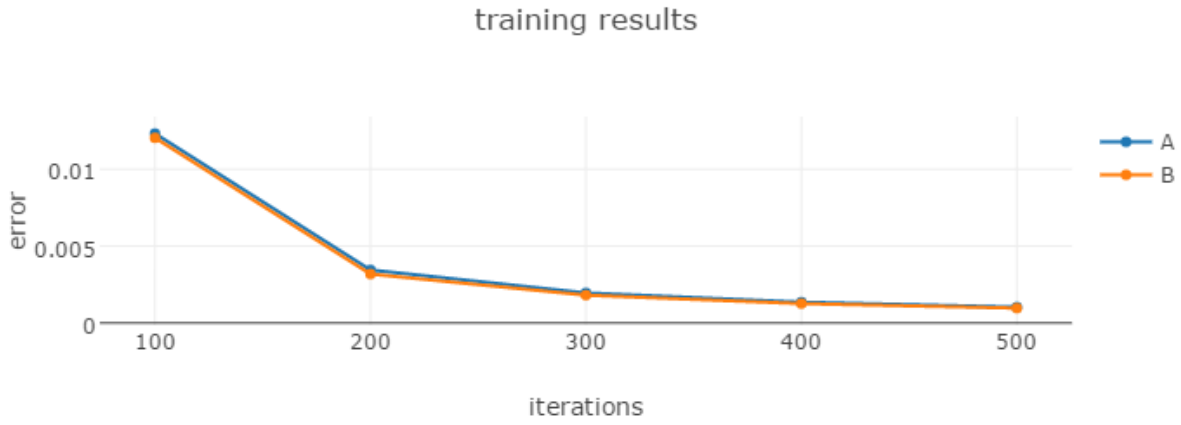

Figura 5.6: Entrenamiento multi-clase 2 gestos: Se puede ver como las 2 redes tiene un comportamiento similar y convergen en etapas tempranas del entrenamiento.

Figura 5.7: Entrenamiento multi-clase 3 gestos: Se puede ver las variaciones en el error para cada una de las muestras usadas en cada red y como cada una encuentra su solución en etapas tempranas del entrenamiento.

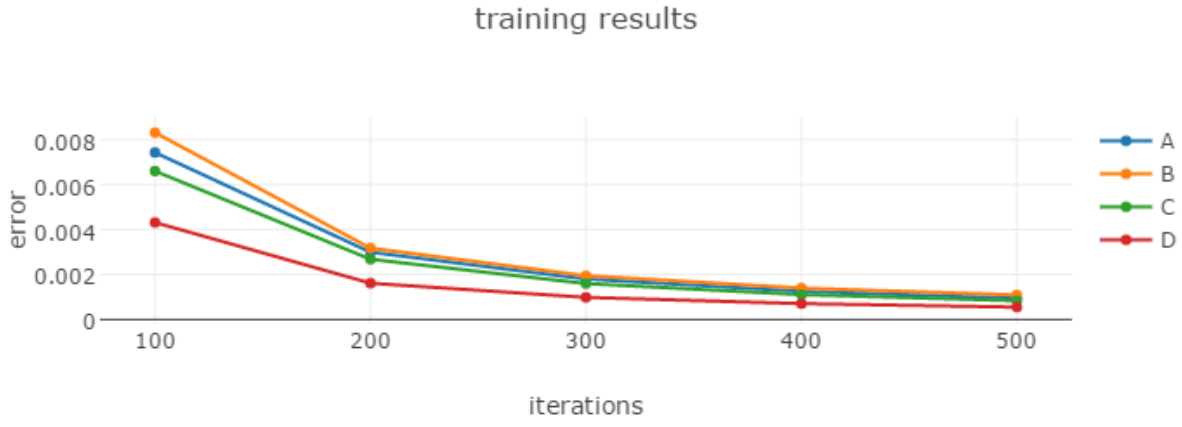

Figura 5.8: Entrenamiento multi-clase 4 gestos: Se puede ver que a pesar de que el numero de gestos y muestras crece el comportamiento de las gráficas es similar para cada una de las redes entrenadas.

De acuerdo a lo visto en las gráficas es importante mencionar que los parámetros para ambos modelos se mantienen estáticos y no se implementó ningún método inteligente para la selección de estos.

La cantidad de iteraciones y tiempo necesarios para tener una red funcional es mucho mayor en el modelo multi-etiqueta que en el modelo multi-clase.

Las comparaciones realizadas se hicieron a partir de los datos obtenidos de las gráficas.

5.2. Gráficas comparativas

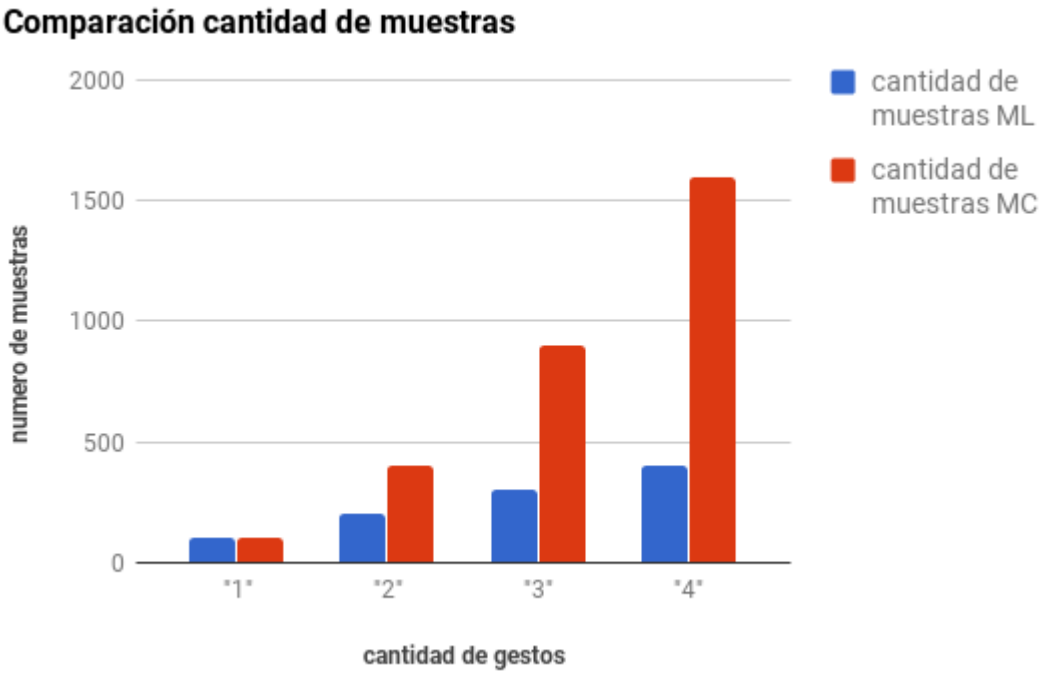

Figura 5.9: Gráfica comparación cantidad de muestras : Se puede observar como el numero de muestras necesarias para el entrenamiento de el modelo multi-clase crece mucho más rápido que en el modelo multi-etiqueta.

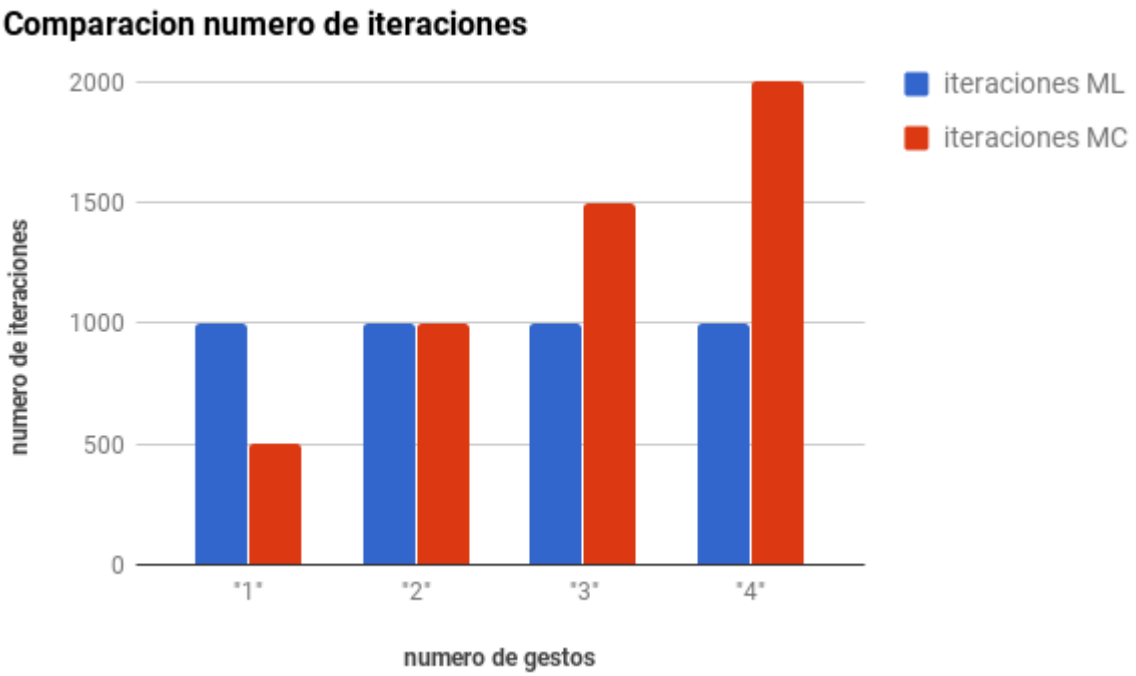

Figura 5.10: Gráfica comparación numero de iteraciones: Se puede observar el numero de iteraciones necesarias para terminar el entrenamiento de ambos modelos según el numero de gestos a clasificar.

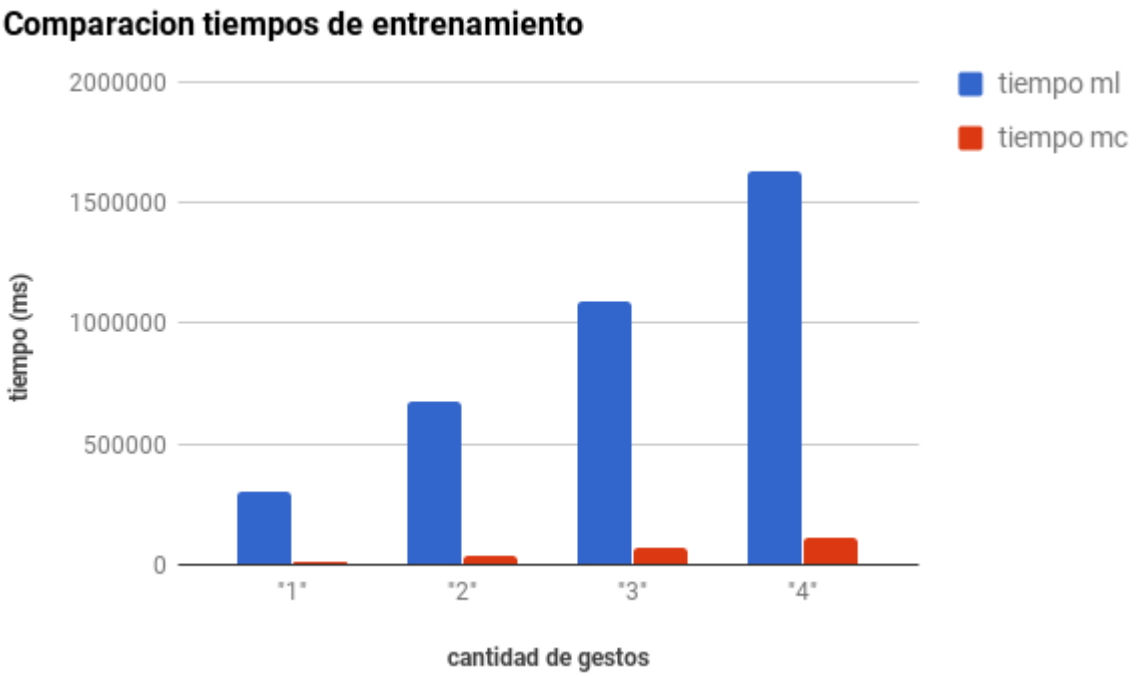

Figura 5.11: Gráfica comparación de tiempos: Se puede observar que el tiempo necesario para llevar a cabo el entrenamiento de la red multi-etiqueta es mucho mayor debido a la cantidad de parámetros involucrados.

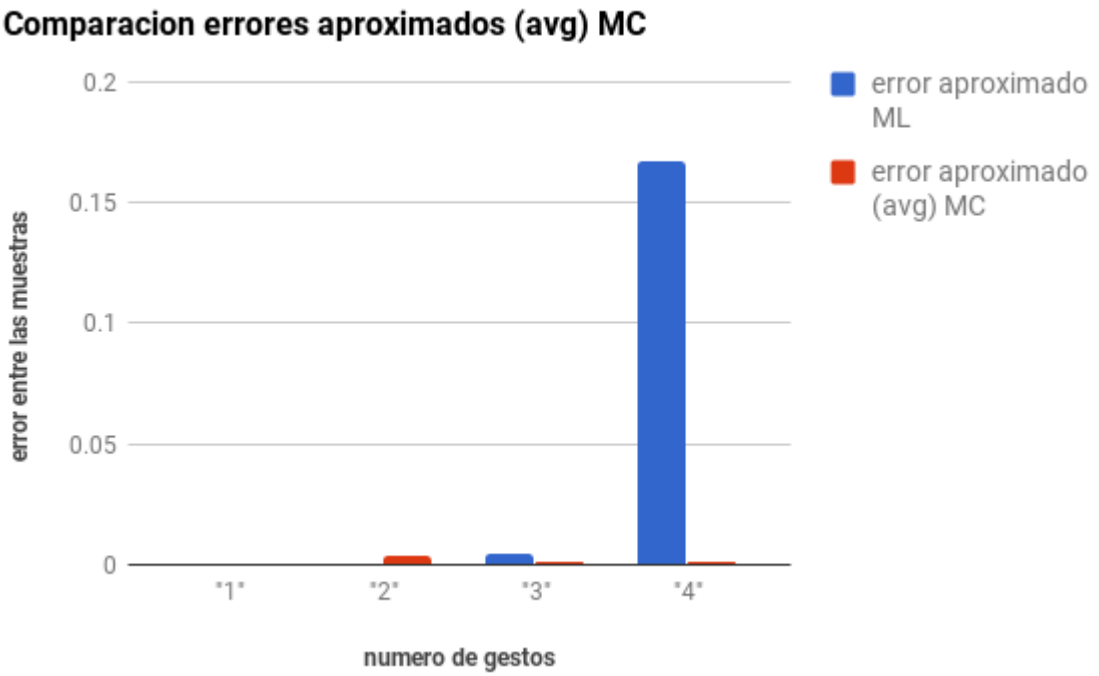

Figura 5.12: Gráfica comparación errores aproximados: Se puede observar como a medida que crece la red se hace notable el error entre las muestras (se hace necesario más tiempo de entrenamiento para llegar a niveles altos de precisión evitando sacrificar la capacidad para generalizar de la red.)

5.3. Aspectos no resueltos

- Es importante mencionar que según las gráficas mostradas anteriormente el numero de interacciones completas no determina si una red esta funcionando adecuadamente, entrenamiento fallido.
- Se debe realizar un numero indeterminado de entrenamientos de la red multi-etiqueta para obtener una red neuronal funcional durante el margen de tiempo estipulado o sera necesario seleccionar otro criterio de parada diferente a 1000 iteraciones, esto debido a que la red neuronal converge en las fases finales del ciclo. A medida que crece el numero de gestos a clasificar y se hace más complejo encontrar una solución (atrapado en un mínimo local o existe una saturación de la capacidad de la red debido a su tamaño fijo).

Figura 5.13: entrenamiento fallido de red con 4 gestos

- No se realizó una medición para el nivel de generalización de las redes debido a la respuesta gráfica de la aplicación.
- No se creo una matriz de confusión para identificar el gesto o gestos con mayor dificultad para reconocer.
- No se implemento ningún algoritmo automático para la selección de parámetros (fading gradient, synaptic prunning).
- No se creo ningún modelo capaz de reconocer gestos dinámicos.
- No se implemento ningún modelo usando SVM(support vector machines).

Realizar esta comparación permite entender un poco sobre los retos de implementar una interfaz basada en gestos del usuario o lo que se conoce como NUI(natural user interfaces) y empezar a incluir estas nuevas tecnologías que permiten capturas de datos dinámicos del usuario para generar soluciones en diferentes campos como la medicina o el entretenimiento.

Capítulo 6

Conclusiones, aportes y recomendaciones

Para el desarrollo del proyecto de clasificación de gestos usando el Leap motion y redes neuronales, se creó una plataforma estable para ayudar en la recolección de datos y entrenamiento de los modelos. La recolección de estas muestras se realizó usando la ayuda de una población de personas que conocen los gestos básicos de la lengua incluyendo algunos de ellos que no poseen ninguna discapacidad auditiva, durante el proceso se tomaron muestras de 10 individuos a los cuales se les pidió que realizaran los gestos del alfabeto dactilológico de la lengua de señas en determinada posición frente a el sensor para realizar una captura apropiada del gesto(1 o 2 gestos por persona) (set de datos creado para primera versión del prototipo), esto resultó ser un proceso arduo debido a la disponibilidad de tiempo de las personas dispuestas a ayudar. Luego de guardar estos resultados en una base de datos, se buscó un número máximo de gestos a "aprender" en un tiempo prudente de entrenamiento(5min - 60min) y a partir de esto se crearon pruebas para los 2 modelos.

Durante este proceso se encontró la utilidad dependiendo del escenario para determinados modelos como identificar diferencias significativas en el tiempo de entrenamiento y la cantidad de datos necesarios para modelos **1 vs N**.

Las comparaciones se hicieron midiendo; tiempos de entrenamiento, error aproximado y cantidad de datos usados para su entrenamiento, para esto se creó un espacio con las herramientas necesarias para validar el funcionamiento de la red usando datos directamente del sensor, luego por medio de gráficas se puede ver el estímulo de ambas redes frente a los datos y una vez se verifique que ha sido entrenada exitosamente la red neuronal, es posible guardarla en una base de datos junto su configuración al final del entrenamiento para ser cargada en otro entorno usando las mismas librerías (synaptic), durante este proceso se determinó que es mucho mejor pensar en otros sensores en el mercado y abarcar un gran espectro de estos a la hora de diseñar un sistema en el futuro.

6.1. Limitaciones del prototipo

- El tipo de gestos que se reconocen solo son estáticos.
- Es importante mencionar que se reconocen gestos de 1 mano a la vez.
- Se usó un solo tipo de sensor: el dispositivo Leap motion para la captura de datos, la literatura habla de obtener mejores resultados integrando más sensores (ejemplo kinect y camaras).
- Se diseñó un aplicativo web como prototipo lo cual implica tener conexión a Internet para acceder a él y su CDN(content delivery network) o un hosting (localhost) e iniciar la base de datos manualmente.
- No está pensado usarse en dispositivos móviles debido a la incompatibilidad con el dispositivo de captura(existen cambios actuales debido a nuevos dispositivos de gama alta y sdks en desarrollo).
- El dispositivo y los métodos usados requieren de altas capacidades de cómputo por lo cual es necesario realizar estas sobre equipos que cumplan con los requerimientos mínimos para hacer uso del hardware y software, además evitar que se vea afectada la experiencia de usuario.
- Es necesario realizar las capturas en un ambiente con iluminación y distancia apropiada para sensor.
- Es necesario hacer capturas en un ambiente libre de oclusión.
- Se recomienda usar características más invariantes como ángulos o valores de las rotaciones respectivas a los ejes coordenados de cada una de las falanges (cosenos directores).
- Se recomienda implementar algoritmos para realizar un ajuste de parámetros automático usando una especie de podado sináptico de acuerdo a los pesos que no aportan información a la solución.

Una de las razones del porqué se genera dificultad al interactuar con personas con discapacidad del habla es el no poder depender de la comunicación verbal para transmitir el mensaje, el problema es bastante grande, así que una solución la cual asista completamente el problema no es viable por medio del prototipo creado en las actuales condiciones, pero teniendo en cuenta nuevas consideraciones en el diseño e incorporando nuevas tecnologías existentes en el mercado, es posible crear soluciones más adecuadas para este tipo de sector, se creo una prueba de concepto usando el prototipo como punto inicial para mostrar su alcance, este proyecto inicialmente no era para el sector de discapacidad auditiva y estaba enfocado principalmente en la exploración de interfaces naturales de usuario para realidad virtual guiadas

por gestos, se decidió aceptar el reto adicional para motivar a las personas a desarrollar para este tipo de sectores usando tecnologías emergentes.

Explorar un poco más sobre interfaces naturales de usuario y su concepto de "gesture control" permite darse cuenta de las posibilidades de interacción en un futuro con los entornos virtuales, ya que esta idea principalmente nació de poder jugar piedra papel o tijera usando periféricos distintos a el mouse, durante la implementación del prototipo funcional mostrando el concepto se decidió buscar un poco más sobre el origen del tema de estudio y se encontró que existen problemas los cuales pueden ser asistidos usando este tipo de tecnologías, en este caso un problema de la población con discapacidad auditiva, con la ayuda de algunas instituciones como: Tecnoparque, la escuela de la palabra y la fundación conectando sentidos, se logro acercar un poco más a la cultura y el tipo de vida que llevan las personas con discapacidad auditiva.

Al ver problemas como la dificultad para aprender una lengua ajena y que su lengua natural no tenga una representación escrita, existen intentos para representarla como lo es la visagrafia, existe la necesidad de abarcar un rango mayor de expresividad como gestos faciales y esto permite darse cuenta de una parte de los requerimientos esenciales de alguna herramienta que planee darle solución al problema completamente.

6.2. Recomendaciones y aportes

- Para entender la Lengua de señas es importante saber que esta es una lengua natural de expresión y configuración gesto-espacial que junto la percepción visual (o táctil si poseen sordo ceguera) permite establecer un canal de comunicación con un entorno social, ya sea conformado por otras personas con discapacidad auditiva o por cualquier otra personas que conozca la lengua de señas. Al igual que un lenguaje oral está sujeto a el proceso universal de cambio lingüístico que hace que surjan variaciones en este a través del tiempo y las regiones, debido a que poseen una estructura gramatical y sintaxis.

- Una característica que ha significado una diferencia entre la lengua de señas y las lenguas orales es la dificultad para ser escrita puesto que se trata de una lengua tradicionalmente ágrafa(no tiene representación escrita) debido a esto la mayoría de personas sordas leen y escriben en la lengua oral de su país. Pese a esto, han habido propuestas para desarrollar sistemas de transcripción de la lengua de señas provenientes de todo el mundo. Un sistema que capture todas las características comunicativas que se utilizan en la lengua de señas mediante signos textuales o bien iconos o ideogramas se han implementado antes para símbolos simples. hamNOSYS, sign writing [12] y le da una posibilidad de tener una representación escrita.

- Algunos países comparten la dactilología pero existen algunos los cuales usan el rostro y el cuerpo o ambas manos para representar un carácter.

- Es necesario motivar a las personas a crear contenido educativo incluyente y encontrar formas de representar un escrito a un medio que permita mostrar todas las capacidades comunicativas de la lengua de señas para facilitar la comprensión de esta.
- Instituto de audiología y su diccionario de Visagrafia ((Lengua visa) representación escrita de la lengua de señas que busca un "Alfabeto (unificado)") (profesor Jaime Hernández). Diccionario de visagrafia físico no muy conocido aún, este fue hecho por el instituto de audiología.
- Se carece de un sistema que interprete el contexto y otorgue el nivel de expresividad necesario.
- Intereses personales sobrepuestos sobre el bienestar de esta población (no se acepta la visagrafia abiertamente aun)
- Gran parte de la lengua depende de lo gestual y la visagrafia no abarca tanta expresividad según interpretes.
- Dificultad extra al tener que aprender otro sistema de símbolos.
- Por ley todos deben ser bilingües y no todos acceden a los beneficios otorgados.
- Requiere una configuración manual y sistemas de referencia además de expresiones faciales.
- Se debe generar una sensibilización hacia los las personas con discapacidad para ofrecer este tipo de tecnologías a este sector y generar soluciones que podrían generar un gran impacto social.

Se plantean a continuación algunas medidas que se pueden tomar para alcanzar los objetivos propuestos en el sistema educativo de cubrir las necesidades de comunicación de la población con discapacidad auditiva:

- Ofrecer las condiciones necesarias para atender a la diversidad en un contexto en el que los niños, puedan desarrollar su potencial sin que se haga énfasis en sus debilidades.
- Diseñar material educativo de calidad para limitar el espectro y enfocarse en el sector educativo en tempranas edades atendiendo todas aquellas necesidades educativas especiales.
- Es necesario personal que conozca de enseñanza de lengua escrita en tempranas edades y como seria el proceso integrado de aprendizaje de las letras y algunas palabras en la lengua de señas
- Temas sencillos como frases y deletreo de palabras con procesos de gamificación podrían ayudar en varios entornos.

Bibliografía

[1] "Leap motion reference." [Online]. Available: https://www.leapmotion.com

[2] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *none*, 1998.

[3] S. kook Jun†, X. Zhou†, D. K. Ramsey‡, and V. N. Krovi†, "A comparative study of human motion capture and computational analysis tools," *none*, 2013.

[4] D. Frankl, "Machine learning in javascript," *none*, 2014.

[5] V. Jakkula, "Tutorial on support vector machine (svm)," *none*, 2012.

[6] "angular.js reference." [Online]. Available: https://angularjs.org

[7] "Synaptic.js reference." [Online]. Available: https://synaptic.juancazala.com/#/

[8] "Leap.js reference." [Online]. Available: https://github.com/leapmotion/leapjs

[9] "Mongodb reference." [Online]. Available: https://www.mongodb.com/es

[10] "Kanban reference." [Online]. Available: https://es.wikipedia.org/wiki/Kanban_(desarrollo)

[11] "Three.js reference." [Online]. Available: https://threejs.org

[12] "Hamnosys, sign writing." [Online]. Available: http://www.signwriting.org/forums/linguistics/ling007.html